Регуляризация в модели линейной регрессии

Напомним, что в общем случае мы получили точное решение для (множественной) линейной регрессии в виде

$$w = (X^T X)^{-1} X^T y$$

Вычислительная сложность аналитического решения — $O(D^2 N + D^3)$, где N — длина выборки, D — число признаков у одного объекта. Слагаемое $D^2 N$ отвечает за сложность перемножения матриц X^T и X, а слагаемое D^3 — за сложность обращения их произведения. Перемножать матрицы $(X^T X)^{-1}$ и X^T не стоит. Гораздо лучше сначала умножить y на X^T , а затем полученный вектор на $(X^T X)^{-1}$: так будет быстрее и, кроме того, не нужно будет хранить матрицу $(X^T X)^{-1} X^T$.

Матрица, стоящая перед вектором y называется псевдообратной матрицей Мура-Пенроуза, это обобщение понятия «обратной матрицы» для неквадратных матриц (мнемоническое правило для запоминания: исходное уравнение $Xw = y$ умножается слева на X^T, после чего появляется квадратная матрица $X^T X$ для обращения). Основные проблемы линейной регрессии связаны с возможной вырожденностью матрицы $X^T X$. Заметим, что эта матрица вырождена, например, когда число объектов в обучающей выборке меньше числа признаков.

Матрица $X^T X$, на практике почти всегда обратима, но зачастую плохо обусловлена. Если признаков много, то между ними может появляться приближённая линейная зависимость, которую мы можем упустить на этапе формулировки задачи. В подобных случаях погрешность нахождения w будет зависеть от квадрата числа обусловленности матрицы X, что очень плохо. Это делает полученное таким образом решение численно неустойчивым: малые возмущения y могут приводить к катастрофическим изменениям w.
Такая ситуация называется **мультиколлинеарностью**. В этом случае у нас, всё равно, возникают проблемы, близкие к описанным выше. Если в выборке два признака будут линейно зависимы (и следовательно, ранг матрицы будет меньше D), то гарантировано найдётся такой вектор весов ν что \nu,x_i⟩ = 0, ∀x_i. В этом случае, если какой-то w является решением оптимизационной задачи, то и $w + a\nu$ тоже является решением для любого a. То есть решение не только не обязано быть уникальным, так ещё может быть сколь угодно большим по модулю. Дело в том, что $X\nu \sim 0$ для вектора ν, состоящего из коэффициентов приближённой линейной зависимости, и, соответственно, $X^T X\nu \approx 0$, то есть матрица $X^T X$ снова будет близка к вырожденной. Как и любая симметричная матрица, она диагонализуется в некотором ортонормированном базисе, и некоторые из собственных значений λ_i близки к нулю. Если вектор $X^T y$ в выражении $(X^T X)^{-1} X^T y$ будет близким к соответствующему собственному вектору, то он будет умножаться на $1/\lambda_i$, что опять же приведёт к

появлению у w очень больших по модулю компонент (при этом w ещё и будет вычислен с большой погрешностью из-за деления на маленькое число).

Важно ещё отметить, что в случае, когда несколько признаков линейно зависимы, веса w_i при них теряют физический смысл. Может даже оказаться, что вес признака, с ростом которого таргет, казалось бы, должен увеличиваться, станет отрицательным. Это делает модель не только неточной, но и принципиально не интерпретируемой. Вообще, неадекватность знаков или величины весов – хорошее указание на мультиколлинеарность.

Есть следующие способы борьбы с вырожденностью:

- регуляризация,
- уменьшение размерности (отбор или выделение признаков),
- увеличение выборки.

Для того, чтобы справиться с этой проблемой, задачу обычно **регуляризуют**, то есть добавляют к ней дополнительное ограничение на вектор весов. Также регуляризацию в общем смысле можно рассматривать как способ уменьшить сложность модели, чтобы предотвратить переобучение или исправить некорректно поставленную задачу. Вместо исходной задачи теперь будем рассматривать следующую:

$$min_w L(f, X, y) = min_w(|Xw - w|_2^2 + \lambda |w|_k^k)$$

где λ это новый параметр, а $|w|_k^k$ принимает одно из двух значений:

$$|w|_2^2 = w_1^2 + \ldots + w_D^2$$

или

$$|w|_1^1 = |w_1| + \ldots + |w_D|$$

Добавочный член $\lambda |w|_k^k$ называется **регуляризационным членом** или **регуляризатором**, а λ – **коэффициентом регуляризации**.
Коэффициент λ является гиперпараметром модели и достаточно сильно влияет на качество итогового решения. Его подбирают по логарифмической шкале (скажем, от `1e-2` до `1e+2`), используя для сравнения моделей с разными значениями λ дополнительную валидационную выборку.

Дополнительно надо отметить, что вес w_0, соответствующий отступу от начала координат (то есть признаку из всех единичек), мы регуляризовать не будем, потому что это не имеет смысла: если даже все значения y равномерно велики, это не должно портить качество обучения.

Ridge регрессия

Регуляризация с помощью $\lambda |w|_2^2$ называется гребневой или Ridge-регрессией (введенная Хоэрлом и Кеннардом в 1970 г), или регуляризацией по Тихонову. По сути

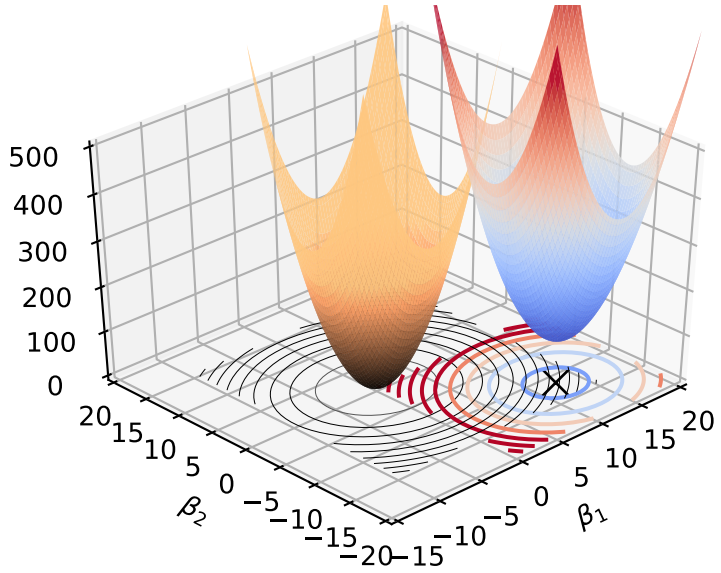
Ридж-регрессия вводит наказание (penalty) за слишком большие коэффициенты. Это наказание представляет собой сумму коэффициентов , возведенных в квадрат (чтобы положительные и отрицательные значения не взаимоудалялись), то есть когда мы видим, что амплитуда значений коэффициентов слишком большая, то попробуем ее уменьшить, добавив ограничение на L^2 норму вектора параметров.

$$\frac{1}{2}\left\|\vec{w}\right\|_2^2 = \frac{1}{2}\sum_{j=1}^{m} w_j^2 = \frac{1}{2}\vec{w}^T\vec{w}$$

Новая функция стоимости примет вид:

$$L_{reg}\left(X, \vec{y}, \vec{w}\right) = \frac{1}{2}\left(\vec{y} - X\vec{w}\right)^T\left(\vec{y} - X\vec{w}\right) + \frac{\lambda}{2}\vec{w}^T\vec{w}$$

Посмотрим на иллюстрацию Ridge - регуляризации.



(Изображение взято отсюда)
При регуляризации у нас появляется, по сути, две функции ошибки: основная (синяя), наказывающая за отклонение истинных значений от прогнозных. и дополнительная (оранжевая), наказывающая за отклонение коэффициентов от нуля. Их сумму и будет пытаться минимизировать алгоритм.

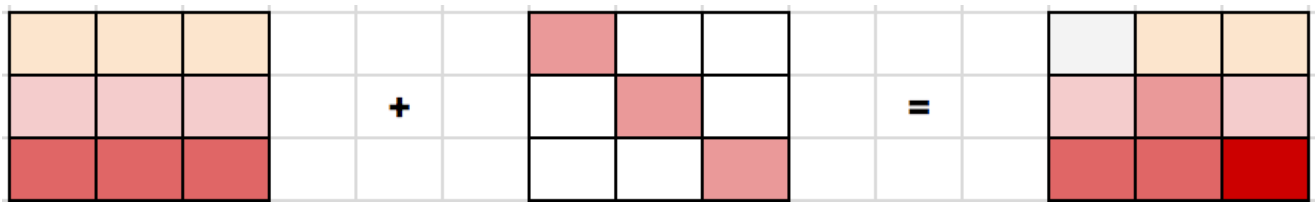
Задача оптимизации при Ridge - регрессии имеет аналитическое решение. Если продифференцировать новую функцию стоимости по параметрам модели, приравнять полученную функцию к нулю и выразить w, то мы получим точное решение задачи.
Вычислим производную по параметрам:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \vec{w}} &= \frac{\partial}{\partial \vec{w}}\left(\frac{1}{2}\left(\vec{y} - X\vec{w}\right)^T\left(\vec{y} - X\vec{w}\right) + \frac{\lambda}{2}\vec{w}^T\vec{w}\right) \\
&= \frac{\partial}{\partial \vec{w}}\left(\frac{1}{2}\left(\vec{y}^T\vec{y} - 2\vec{y}^T X\vec{w} + \vec{w}^T X^T X\vec{w}\right) + \frac{\lambda}{2}\vec{w}^T\vec{w}\right) \\
&= -X^T\vec{y} + X^T X\vec{w} + \lambda\vec{w}
\end{aligned}$$

И найдем решение в явном виде:

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \vec{w}} = 0 \quad &\Leftrightarrow \quad -X^T\vec{y} + X^T X\vec{w} + \lambda\vec{w} = 0 \\ &\Leftrightarrow \quad X^T X\vec{w} + \lambda\vec{w} = X^T\vec{y} \\ &\Leftrightarrow \quad \left(X^T X + \lambda E\right)\vec{w} = X^T\vec{y} \\ &\Leftrightarrow \quad \vec{w} = \left(X^T X + \lambda E\right)^{-1} X^T\vec{y}\end{aligned}$$

здесь E — единичная диагональная матрица. Такая регрессия (с регуляризованной функцией стоимости) называется гребневой регрессией, где гребнем является как раз диагональная матрица, которую мы прибавляем к матрице $X^T X$, в результате чего гарантированно получается регулярная матрица.



В свою очередь, градиент функции потерь по весам (для градиентного спуска) выглядит так:

$$\nabla_w L(f, X, y) = 2X^T(Xw - y) + 2\lambda w$$

Таким образом мы получаем обновлённую версию приближенного алгоритма, отличающуюся от старой только наличием дополнительного слагаемого.

LASSO регрессия

Лассо́-регре́ссия (англ. lasso или LASSO, least absolute shrinkage and selection operator — оператор наименьшего абсолютного сокращения и выбора), называется разновидность линейной регрессии, которая используется для решения проблемы мультиколлинеарности и отбора наиболее информативных (с точки зрения способности объяснять дисперсию зависимой переменной) признаков.
Попробуем теперь ограничить вектор параметров модели, используя L^1 норму:

$$\left\|\vec{w}\right\|_1 = \sum_{j=1}^{m} |w_j|$$

Тогда задача примет вид:

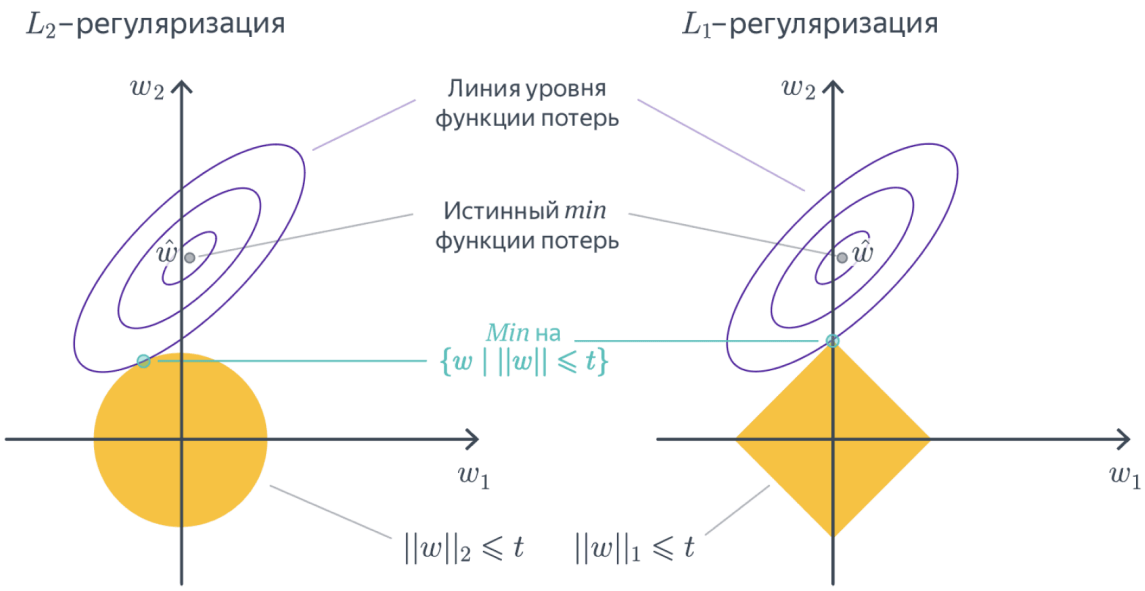
$$L_{reg}\left(X, \vec{y}, \vec{w}\right) = \frac{1}{2n}\sum_{i=1}^{n}\left(\vec{x_i}^T\vec{w} - y_i\right)^2 + \lambda\sum_{j=1}^{m}|w_j|$$

L^1 - регуляризация приводит к тому, что коэффициенты модели у наименее информативных признаков приравниваются к нулю. Это позволяет проводить отбор признаков в модели и сделать модель более интерпретируемой. Метод был предложен Р. Тибширани в 1996 г.

Таким образом при L^1 - регуляризации мы можем удалять признаки, слабо влияющие на таргет. Это также даёт возможность автоматически избавляться от признаков,

которые участвуют в соотношениях приближённой линейной зависимости, соответственно, спасает от проблем, связанных с мультиколлинеарностью, которые обсуждались выше.

Давайте рассмотрим следующее изображение (классическую визуализацию границ двух видов регуляризации из книги T. Hastie *The Elements of Statistical Learning*)



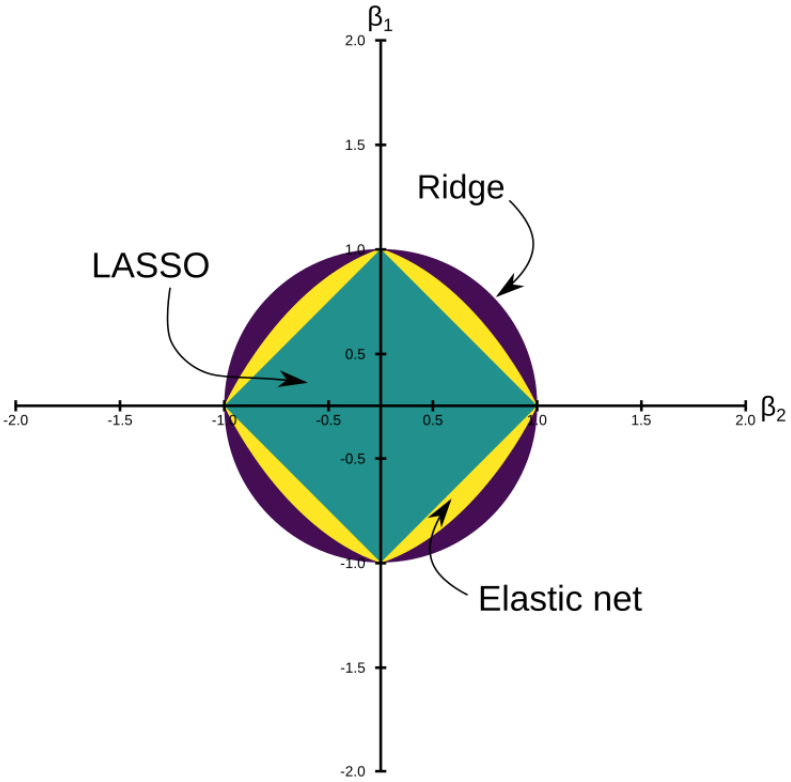
Линии уровня L_1 - нормы – это N - мерные октаэдры. Точки их касания с линиями уровня лосса, скорее всего, лежат на грани размерности, меньшей $N-1$, то есть как раз в области, где часть координат равна нулю.

Elastic Net регрессия

Регрессия Elastic Net использует как L^1, так и L^2 регуляризацию.

$$min_w(|Xw - w|_2^2 + \lambda_1|w|_1^1 + \lambda_2|w|_2^2)$$

В процедуре регуляризации с помощью эластичной сетки сначала мы находим коэффициент гребневой регрессии. После этого мы выполним алгоритм лассо для коэффициента регрессии ridge, чтобы уменьшить коэффициент.



Здесь мы можем видеть, что после выполнения регрессии ridge регрессия lasso участвует в процедуре, которая учитывает все переменные из набора данных.

ElasticNet был впервые представлен Цзоу и Хасти в 2005 г. Этот метод был разработан как расширение регрессии по гребню и лассо, устраняющее ограничения обоих методов. Гребневая регрессия использует штраф L_2 для уменьшения коэффициентов, но не выполняет выбор переменной. С другой стороны, регрессия Лассо использует штраф L_1, который приводит к разреженным решениям за счет установки некоторых коэффициентов равными нулю. Однако Lasso борется с коррелированными предикторами и имеет тенденцию выбирать только одну переменную из группы высококоррелированных переменных.

1. Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(2), 301-320. URL
2. Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. Technometrics, 12(1), 55-67. URL
3. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1), 267-288. URL